Available online at www.sciencedirect.com

SCIENCE DIRECT®

Data & Knowledge Engineering 53 (2005) 75–97

DATA & KNOWLEDGE ENGINEERING

www.elsevier.com/locate/datak

Biological workflow with BlastQuest

William G. Farmerie [a,*], Joachim Hammer [b], Li Liu [a], Anuj Sahni [a], Markus Schneider [b]

[a] *ICBR Molecular Services Division: DNA Sequencing Core, Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL 32611, USA*

[b] *Department of Computer and Information Science and Engineering*

Received 22 June 2004; accepted 22 June 2004
Available online 3 August 2004

Abstract

Besides domain-specific biological problems, biologists are confronted with many computational problems. The large amount of varying, heterogeneous, and semi-structured biological data, the increasing complexity of biological applications, methods, and tools afflicted with uncertainty and missing knowledge, as well as the lacking interoperability of available tools necessitate integrative measures to enable biology workflow. In this paper we address these problems in the context of the processing and evaluation of BLAST query results. We present a new tool, called *BlastQuest*, which relies on database technology and provides sophisticated interactive and Web-enabled query, analysis, and visualization facilities for genomics data. The interface with the Gene Ontology and the KEGG pathway databases decisively foster the biological workflow. Finally, based on our experience with BlastQuest, we briefly sketch a new concept, called *Genomics Algebra*, for solving genomic data management problems from a broader perspective.

Keywords: BLAST; Gene Ontology; KEGG pathway database; Genomics Algebra; Unifying database

[*] Corresponding author.
E-mail address: wgf@biotech.ufl.edu (W.G. Farmerie).

0169-023X/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2004.06.014

1. Introduction

Besides domain-specific biological problems, biologists are confronted with many computational problems. For example, the exponentially growing volume of heterogeneous, semi-structured biological data that has to be processed and analyzed. Another problem is the increasing complexity of biological applications, methods, and tools afflicted with an inherent lack of biological knowledge as well as intrinsic uncertainty. A third problem is the lacking interoperability of available tools, i.e., biological tools are more or less self-contained and isolated, mostly only visualization-based, and incapable of exchanging data with each other. As a result, many challenges in biology and genomics are now challenges in computing and here especially in information management and algorithm design. This necessitates the development of an appropriate "communication interface" between biologists and computer scientists who each live in their own world but also recognize the chances for jointly solving important problems in common future research.

This paper deals with the three aforementioned problems in the context of BLAST (Basic Local Alignment Search Tool) [1], a common tool for conducting similarity searches. BLAST comprises a set of similarity search algorithms that employ heuristics to detect relationships between gene sequences and that rank the computed "hits" statistically. An essential problem for the biologist is currently the processing and evaluation of BLAST query results, since a BLAST search yields its result exclusively in a textual format (e.g., ASCII, HTML, XML). This format has the benefit of being application-neutral but at the same time prevents efficient analysis.

In this paper, we describe a new powerful tool called *BlastQuest* for managing BLAST results stemming from multiple individual queries. This tool provides the biologist with interactive and Web-enabled query, analysis, and visualization facilities beyond what is possible by current BLAST interfaces. In particular, BLAST results from multiple queries are imported, structured, and stored in a relational database to support a series of built-in analysis operations that can be used to select, browse, filter, group, order, search, and annotate sequence data efficiently and without referring to the original BLAST result files. In addition, users have the option to interact with the data through a forms-based query interface. BlastQuest also establishes connections to the *Gene Ontology* (GO) [10], which is a controlled vocabulary about gene and protein roles in cells, and the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) [16], which is a pathway database and integrates current knowledge on molecular interaction networks in biological processes.

The rest of this paper is organized as follows. Section 2 briefly reviews some important biological concepts and addresses the biological workflow when working with BLAST, Gene Ontology, and the KEGG database. Section 3 emphasizes the need for tools capable of processing BLAST results and identifies their functional requirements. In Section 4, we describe our BlastQuest prototype from the system architecture and implementation perspective. An example session in Section 5 describes the main features and data analysis options of the BlastQuest system and its user interface. Section 6 evaluates the BlastQuest system and reports on our experiences with it. Section 7 discusses related work. Section 8 considers desired improvements to BlastQuest. We conclude the section with a brief description of a new, data model, language, and architecture called *Genomics Algebra* for integrating, processing and querying genomic information. Finally, Section 9 summarizes the paper.